Bengali Printed Character Recognition – A New Approach

Soharab Hossain Shaikh[1], Marek Tabedzki[2], Nabendu Chaki[3], and Khalid Saeed[4]

[1] A.K.Choudhury School of Information Technology, University of Calcutta, India
soharab.hossain@gmail.com
[2] Faculty of Computer Science, Bialystok University of Technology, Poland
m.tabedzki@pb.edu.pl
[3] Department of Computer Science & Engineering, University of Calcutta, India
nabendu@ieee.org
[4] Faculty of Physics and Applied Computer Science,
AGH University of Science and Technology, Cracow, Poland
saeed@agh.edu.pl

Abstract. This paper presents a new method for Bengali character recognition based on view-based approach. Both the top-bottom and the lateral view-based approaches have been considered. A layer-based methodology in modification of the basic view-based processing has been proposed. This facilitates handling of unequal logical partitions. The document image is acquired and segmented to extract out the text lines, words, and letters. The whole image of the individual characters is taken as the input to the system. The character image is put into a bounding box and resized whenever necessary. The view-based approach is applied on the resultant image and the characteristic points are extracted from the views after some preprocessing. These points are then used to form a feature vector that represents the given character as a descriptor. The feature vectors have been classified with the aid of k-NN classifier using Dynamic Time Warping (DTW) as a distance measure. A small dataset of some of the compound characters has also been considered for recognition. The promising results obtained so far encourage the authors for further work on handwritten Bengali scripts.

Keywords: Bengali character, view-based algorithm, layer-based method, bounding box, unequal partition.

1 Introduction

Character recognition has been a popular field of research for past few decades. Research, in this arena, has been done not only on Bengali but also on some other languages [3], [4]. In [3], a method for handwriting recognition is proposed for Polish alphabet. It is based on Toeplitz matrix minimal Eigen values approach. In [4] a Template Matching based signature recognition algorithm is presented. In [11] a successful trial was made to recognize both typewritten and handwritten English and Arabic texts without thinning on the basis of region growing segmentation. In this

K. Saeed et al. (Eds.): CISIM 2013, LNCS 8104, pp. 129–140, 2013.

work, however, and following the view-based approach of [5], [6], the Bengali language is studied for automatic recognition. Recognition of Bengali script has a lot of importance. Bengali is one of the most popular languages in India. All over the world more than 200 million people speak in Bengali and this is the second most popular script next to Devanagari in India. It also suggests the scripts of two other languages, Assamese and Manipuri. Bengali is the official language of Bangladesh, a neighbour of India.

Recognition of Bengali printed as well as handwritten characters has been a popular area if research in the arena of OCR for past few years as found in the literature [1, 2, 6, 9, 13, 14, 15]. Research is being done on the recognition of both the basic [10] and compound [9] Bengali characters. Attempts have also been made in the recognition of Bengali numerals [13], [18]. The modern Bengali alphabet set consists of 11 vowels and 39 consonants. These characters are called *basic characters*. Bengali text is written from left to right. The concept of upper/lower case is missing in Bengali. Most of the Bengali characters have a running horizontal line on the upper part of the characters; this line is known as *Matra*.

Characters in Bengali are not alphabetical as in English (or Roman) where the characters largely have one-sound one-symbol characteristics. It is a mixture of syllabic and alphabetic characters [9]. The use of modified and compound characters is also very common in Bengali. This paper presents methods for recognizing Bengali printed characters based on view-based approach. Both the top-bottom and left-right view-based approaches have been considered. This work is an extension of [6]. In this paper we have considered unequal partitions of the character images. Also a set of compound characters have been considered for view-based analysis.

The rest of the paper is organized as follows: section 2 is a short review of the existing literature. Section 3 describes the major functional steps involved in the recognition process and feature extraction methods. In section 4 the concept of unequal partitioning is presented followed by the considerations for compound characters. Classification and experimental results are given in section 5.

2 Previous Work

Different techniques have been found in the literature for optical character recognition. The curvelet transform has been heavily utilized in various areas of image processing. In [10] a novel feature extraction scheme is proposed on the basis of the digital curvelet transform. The curvelet coefficients of an original image as well as its morphologically altered versions are used to train separate k–nearest neighbour classifiers. Output values of these classifiers are fused using a simple majority voting scheme to arrive at a final decision. In [22] a method has been suggested based on curvature-based feature extraction strategy for both printed and handwritten Bengali characters. BAM (Bidirectional Associative Memories) neural network has been used in [19] for Bengali character recognition. The conventional methods are used for text scanning to segmentation of a text line to a single character. An efficient procedure is proposed for boundary extraction, scaling of a character and the BAM neural network which increases the performance of character recognition are used. In [20] a modified

learning approach, using neural network learning for recognizing Bengali characters, has been presented. Research has been done on the recognition of handwritten Bengali characters [14]. Multi-Layer Perceptron (MLP) trained by back-propagation (BP) algorithm have been used as classifier.

In [18] an automatic recognition scheme for handwritten Bengali numerals using neural network models has been presented. A Topology Adaptive Self Organizing Neural Network is first used to extract from a numeral pattern a skeletal shape that is represented as a graph. Certain features like loops, junctions etc. present in the graph are considered to classify a numeral into a smaller group. If the group is a singleton, the recognition is done. Otherwise, multilayer perceptron networks are used to classify different numerals uniquely. Hidden Markov Models (HMMs) are used for both online and offline character recognition systems for different scripts around the world. A OCR program that uses HMM, for recognition process, has been made for Bengali documents in [12]. For using HMM it is required to have a sequence of objects to traverse through the state sequence of HMM. So the features are shaped into a sequence of objects. For each character component, a tree of features is made and finally the prefix notation of the tree is applied to the HMM. In the tree, the number of child of a node is not fixed, so, the child-sibling approach is applied to make the tree. Hence the prefix notation of the tree will contain nodes in the order: root, prefix notation of the tree rooted at its child, prefix notation of the trees rooted at the child's siblings from left to right order. After that HMM is used for the recognition purpose. Attempts have also been made on methods of segmentation and recognition of unconstrained offline Bengali handwritten numerals [13]. A projection profile based heuristic technique is used to segment handwritten numerals. A neural network based classifier is used for classification purpose. Paper [23] addresses various aspects of the problems associated with processing and recognition of printed and handwritten Bengali numerals. A scheme is proposed in this work for recognizing handwritten as well as printed numerals with different fonts and writing styles including noisy and occluded numerals. Polygon approximation is used to represent the contours of the letters. After that Fourier descriptors are used as shape features. The standard Multi-Layer Perceptron (MLP) augmented with MAXNET was used as a classifier. In [21] a method has been presented based on primitive analysis with template matching to detect compound Bengali characters. Most of the works on Bengali character are recognition of isolated characters. Very few papers deal with a complete OCR for printed document in Bengali. In [17] a chain code method of image representation is used. Thinning of the character image is needless when chain code representation is used. The main difficulties in printed Bengali text recognition are the separation of lines, words and individual characters. In [16] a new approach has been proposed to segment and recognize printed Bengali text using characteristic functions and Hamming network. A new algorithm has been proposed to detect and separate text lines, words and characters from printed Bengali text. The algorithm uses a set of characteristic functions for segmenting upper portion of some characters and characters that come under the Base line. It also uses a combination of Flood-fill and Boundary-fill algorithm for segmenting some characters that cannot be segmented using traditional approach. Hamming network is used for recognition scheme.

Recognition is done for both isolated and continuous size independent printed characters. In [15] a study has been made on handwritten Bengali numerals.

3 Major Functional Steps

Figure 1 shows the flowchart of major functional steps which have been outlined as follows:

i) Binarization: Printed documents written in Bengali have been scanned using a flat-bed scanner. Samples have also been collected using software supporting different Bengali fonts. These samples are converted to images and all the samples have been binarized.

ii) Segmentation: The documents contain Bengali text. Individual character has to be extracted from the text before applying view-based approach. Histogram of individual pixel row and columns of the text is computed. The individual lines containing many words have been segmented out from the text image using a horizontal histogram. The individual letters have been segmented out from the images of lines of text using vertical histogram.

iii) Matra Removal: The *Matra* is removed from top of the character. Standard image-editing software is used for doing the same. After removing the *Matra*, the characters without Matra is stored. View based approach is performed on these images. The importance of this phase is detailed out in section 3.1.1.

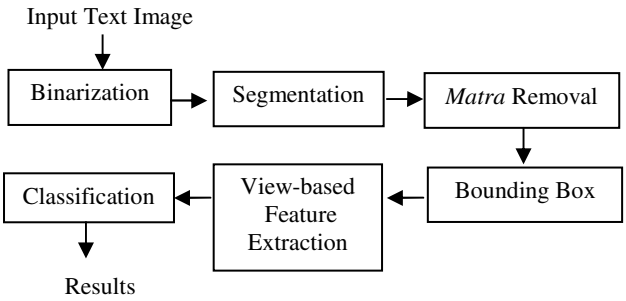

Fig. 1. Flow-chart of Major Functional Steps

iv) Applying Bounding Box: The character is put into a bounding box (rectangle that most tightly contains the character) before applying the view-based approach. The bounding box may be used as an indicator of the relative positions of features in a character.

v) View-based Feature Extraction: The features are extracted from four views of each individual letter. Additionally, the number of changes of the pixel values from white-to-black and vice versa have been calculated for each row and column. In inner-views approach the views of partitioned image are used to extract the features. These values form the feature vector representing the particular letter. This is detailed

out in section 3.1. This vector is finally fed to the classifier for classification and recognition.

vi) Classification: The feature vectors are classified with the help of Dynamic Time Warping distance, k-NN and Ensemble of Classifiers in the present experiment. This has been explained in section 5.

3.1 Methods for Feature Extraction

In this section the view-based approach is detailed out. Also the layer based approach and the concept of bounding box is described.

3.1.1 View-Based Approach

This idea was presented in [5], [6] and [8]. The method is based on the fact that for correct character recognition one usually needs only partial information about its shape – or contour.

The "views" of given object are examined to extract from them a characteristic vector, which describes its shape. The view is representation of the coordinates of the points lying on one of four boundaries of the object (top, bottom, left or right) – it consists of pixels belonging to the contour of a word and having extreme values of x (for left and right view) or y coordinate (for top and bottom). The conventional preprocessing stage of thinning is not essential here because in this approach only the shape of the character is analyzed. However, and in some cases, as will be seen later, the thinning process helps improve the results. The binarization is required to convert the scanned image into a black-and-white one. The main feature extraction step is to take out all the characteristic points from the views.

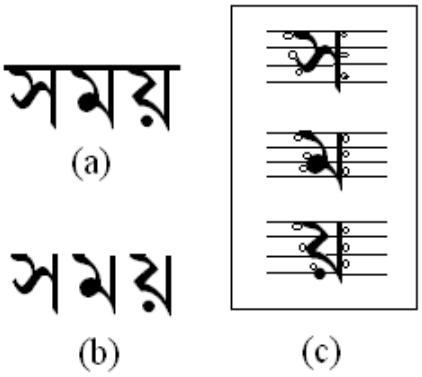

Fig. 2. (a) The Bengali word *Somoy* with *Matra*. (b) *Somoy* after removing *Matra*, (c) The letters of the word segmented and 3 horizontal partition is done. Also characteristic points have been marked on the left and right contour in each segment for Lateral view based consideration.

Thus, characteristic vector for top view consists of maximal values from each pixel-column on an image. To reduce dimensionality of obtained vectors one can introduce some form of down-sampling. In our case it is equivalent to searching for characteristic points in each view that describes the shape. In [7] several methods of selecting these points have been explored. In the method used in this paper, characteristic values are found by dividing original vectors into *n* identical segments, and calculating mean value for each one. In this way four characteristic vectors describing the given letter is found. Next, these four vectors are transformed into a one vector, which describes the given letter.

The top-bottom view based approach cannot be directly applied to Bengali characters. This is because most of the Bengali characters have a line known as the "*Matra*" on top of them. So, the top view for almost every character (except a few characters that do not have a *Matra* on top of them) is same. So the *Matra* should be removed first; after that the character without *Matra* can be processed for feature extraction using top-bottom view-based approach. *Matra* removal phase is not necessary if only lateral-view based approach is applied for feature extraction. But here both the lateral as well as top-bottom view based approaches have been applied, so removing Matra is necessary.

Figure 2 presents this process for a word "Somoy" meaning time in Bengali.

3.1.2 Layer-Based Approach

In layer based approach the number of changes of pixel values from black to white (or transitions from foreground to background) is considered. It tells whether the pixel-column (or row) contains one or several strokes, as is presented in Figure 3. This way some information about the *inner shape* of the letters can be obtained. For example two Bengali characters *Ja* and *Sha* have been considered here. Both the characters look exactly the same as far as the contour is concerned. The only difference is in the internal shape of the characters. Layer based approach can be applied to distinguish these letters. The same can be done for pixel rows. With the use of layer-counting (layer-view), improvements in results have been achieved for word recognition as in [15].

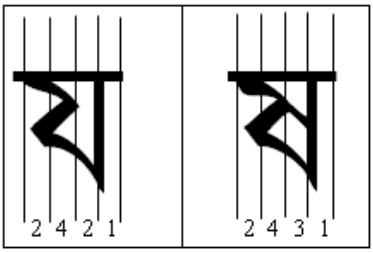

Fig. 3. Number of transitions for the letters *Ja* and *Sh*

Two variations of this approach have been considered. In each pixel column the number of changes of colour from white to black and black to white is counted. The same approach is performed for the pixel rows.

When considered the logical partitioning of the image (3 and 6 partitions) the average number of colour changes in each partition is calculated and used to form a feature vector. However, that lead to very poor recognition accuracy. So, later the whole image without any partition is considered and the layer-based approach over all the pixel rows and columns of the binary image has been performed.

3.1.3 Inner-View Approach

In order to improve the effectiveness of detection another idea was proposed. It implies the inclusion of additional information in the feature vector about the internal structure of the characters. It took the form of internal views. This is an innovative proposal used for the first time.

As with the classical view-based approach, there are also four possible views – top, bottom, left and right. This time, however, they do not form the outer shape of the letter, but the internal one. For this purpose, the character image is divided into two parts. For example, to obtain the top and bottom views, the image is divided horizontally into lower and upper parts. Then a bottom view of the upper half and a top view of the lower one can be obtained. Similarly, for the right and left view - the internal view of the left and right halves is created. The method of characteristic vector extraction for each view does not differ from the one presented in section 3.1.1.

4 Handling Compound Characters

A compound character is one that is formed by clubbing more than one basic character. Examples of some compound characters in Bengali are given in Figure 4. The basic characters forming those compound characters are also shown. There are a few problems in recognizing the compound characters.

Firstly, the problem is that there is no standard segmentation method that suits for all the compound characters properly. Component basic characters may retain their shapes in the compound character. Figure 4 shows two groups of compound characters. In the first group, the shapes of the component basic characters are retained in the compound one whilst in the second group, the basic shapes of the component characters are not retained.

So if the compound character is segmented properly then the basic component characters can be extracted out of it and on the basis of information of the component characters, the compound character can be identified. But as shown in Figure 4, first group, that no specific segmentation method applies well for all the compound characters. For the characters at line 1 and 2 a horizontal segmentation works well. For the characters at line 3 and 4, a diagonal segmentation is required, whereas for the character at line 4, a horizontal segmentation will be suitable.

Secondly, as shown in the second group of Figure 4, the shape of the compound characters may change significantly when two basic characters are clubbed to form it. This compound character contains no clue of the basic components. So no segmentation can be done for recognizing such characters.

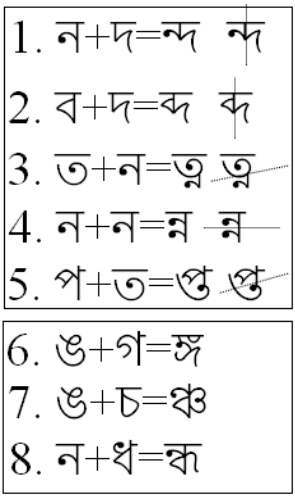

Fig. 4. Compound characters and segmentation for extracting the individual basic component characters

Thus, the methods based on the segmentation of compound characters do not give satisfactory results for recognizing compound characters. In this paper view-based approach is applied for performing this recognition task. The idea is that the contour of each compound character is unique. So, applying view-based approach a feature vector can be formed for each of the compound characters. This feature vectors should be stored in a vector codebook database. Whenever a character has to be identified, the view-based approach is applied to it to form a feature vector. This vector is then compared with those at the codebook to find a similar match. This way a compound character can be identified.

5 Classification and Experimental Results

For the purpose of classification the method based on k-Nearest Neighbor (k-NN) algorithm has been used. It is a type of instance-based learning, where classification is based on closest training examples in the feature space. The training set is composed of labeled examples (feature vectors of correctly classified Bengali characters). In the classification phase, an unlabeled vector is classified by assigning the label which is most frequent among the k training samples nearest to that vector. In the present experiments distance is calculated using DTW approach [4], [24].

In all experiments K-fold cross-validation method is used, so original database was randomly partitioned into K (where $K = 10$) equal size subsets. Classification process was then repeated k times, with each of the k subsets used exactly once as a test set, where the remaining K -1 subsets are used as training data. The results from K folds were then combined to produce recognition result estimation.

Tests were performed on the collected database of 1000 characters. Images were binarized, cleaned of noise and prepared to recognize.

Table 1. Results of classification

Test	Views	Layers	Inner Views	KMM	Features	Recognition rate
1	✔				32	43.7%
2	✔			✔	32	57.2%
3	✔			✔	40	56.5%
4	✔			✔	24	55.9%
5		✔		✔	16	46.2%
6			✔	✔	32	62.7%
7	✔	✔	✔	✔	80	**76.8%**

Authors performed a series of experiments, designed to allow the determination of the optimal strategy and parameters. The following table shows some sample results.

In the first approach only views were used (four for each character), wherein from each view 8 points were collected giving a 32-element long feature vector. This has resulted in a rather low efficiency of about 44% characters correctly classified.

The next experiment revealed that prior character skeletonization allows for a significant improvement in efficiency. For this purpose authors used KMM algorithm [25]. In this approach the character image was thinned to a skeleton form before applying view-based method to extract features. Thanks to this the percentage increased to 57%.

Approach 3 and 4 show an attempt to determine the optimal number of points extracted from each view. As can be seen an attempt to reduce or increase this value does not bring improvement indicating that for the studied characters, the optimal number proved to be eight points per view.

The fifth study tested the effectiveness of the layer-based approach. For this purpose the algorithm of section 3.1.2 was used, designating the number of transitions for the vertical and horizontal directions. For each of the two directions only 8 values were considered – using the very same method of downsampling, as in the view-based approach described in section 3.1.1. The percentage is lower than that of the view-based approach, but the attention should be paid to the smaller size of the vector as well as that both of these approaches will be used jointly.

Experiment 6 shows a novel approach used for the first time as presented in this work - it is based on the use of internal views as described in section 3.1.3. Also four views were used and 8 characteristic points were collected from each view. As can be seen, this allows for a larger percentage than the classical approach of over 60%.

The last row of the table shows the results obtained by a combination of all these approaches. In this manner, the effectiveness of recognition has reached 76.8%. The cost is obtaining a large feature vector due to the fact that all of the features of the previous studies: views, layers and inner views were used. The solution to this issue would be to carry out a prior feature selection, it would help to determine which of features are relevant to the process of classification and which may be removed.

Table 2 collects the results of the last survey (approach no 7) breaking them into some selected characters. This allows to observe the individual results of each of

them. Columns of the table presents the selected letter and the results of recognition. As can be seen, some characters were classified with efficiency exceeding 90%, while some decreases the performance.

Table 2. Results of classification for individual letters

Letter	Recognition rate
ঞ	95%
ঢ়	90%
ড়	85%
ঙ	70%
গ	70%
ন	60%
ম	50%
ঘ	50%
প	40%
খ	35%
স	25%
Overall	**76.8%**

In order to determine the causes of low rate obtained for some of the characters, the results of recognition were formed into matrix of confusion, to expose most common mistakes. This revealed that some of the characters were particularly often confused with each other. Some of these pairs are shown in Table 3.

Table 3. Most common mistakes

Pairs of confused letters
ম – স
গ – প
ঘ – ঘ
খ – খ

Each of these pairs was confused with each other in at least 20% of cases. As can be seen they are actually very similar to each other and of very similar shape. Hence, the algorithm could not distinguish between them.. Further work will be aimed at more closely look at the similar pairs of characters to find specifications that allow them to differentiate and thus improve the efficiency of recognition.

6 Conclusions and Future Work

In this paper, the authors have proposed a new method for Bengali printed character recognition using View-based and Layer-based methods. The concept of unequal partitioning has been applied in this study. A classifier based on DTW and k-NN has been used for the experimentation purposes. Recognition accuracy of 76.8% has been achieved which is considered to be very high as the Bengali alphabets resemble handwriting rather than type writing. It contains alphabets that are cursive in nature. The proposed method therefore, will be the basis of handwritten character recognition.

The view-based approach is being applied for Bengali character recognition for the first time. The experiments show gradual improvements. The future plan is to modify the recognition engine and use a different pattern classifier like a neural network to study the performance of the proposed method.

Acknowledgement. This work was supported by Bialystok University of Technology grant S/WI/2/2013.

References

[1] Bag, S., Bhowmick, P., Harit, G.: Recognition of Bengali Handwritten Characters Using Skeletal Convexity and Dynamic Programming. In: 2nd International Conference on Emerging Applications of Information Technology (EAIT), Kolkata, pp. 265–268 (2011)

[2] Naser, M.A., Hossain, M.M., Tito, S.R., Hoque, M.A.: Recognition of Bangla Characters using Regional Features and Principal Component Analysis. In: International Conference on Electrical and Computer Engineering (ICECE), Dhaka, pp. 506–509 (2010)

[3] Tabędzki, M.: Handwriting Recognition System on the basis of Toeplitz Matrices. PhD Thesis, Bialystok University of Technology, Bialystok, Poland (2010)

[4] Adamski, M.: Signature Recognition by the use of Template Matching Techniques. PhD Thesis, Bilystok University of Technology, Bialystok, Poland (2010) (in Polish)

[5] Tabedzki, M., Rybnik, M., Saeed, K.: New Results for View-Based Feature Extraction Method for Handwritten Words Recognition without Segmentation. In: 1st International Conference on Image Processing & Communications, Poland (2009)

[6] Shaikh, S.H., Chaki, N., Tabedzki, M., Saeed, K.: A View-based Approach for Recognition of Bengali Printed Characters. In: 8th International Conference on Computer Information Systems and Industrial Management (CISIM), Coimbatore, India (December 2009)

[7] Tabedzki, M., Rybnik, M., Saeed, K.: A Method for Handwritten Word Recognition without Segmentation. Polish Journal of Environmental Studies 17(4C), 47–52 (2008)

[8] Saeed, K., Tabedzki, M.: New Experiments on Word Recognition Without Segmentation. In: Advances in Information Processing and Protection, Part-III, pp. 323–332. Springer, US (2008)

[9] Pal, U., Wakabayashi, T., Kimura, F.: Handwritten Bangla Compound Character Recognition using Gradient Feature. In: 10th International Conference on Information Technology (ICIT), pp. 208–213 (December 2007)

[10] Majumdar, A.: Bengali Basic Character recognition using Digital Curvelet Transform. Journal of Pattern Recognition Research, 17–26 (2007)
[11] Saeed, K., AlBakoor, M.: Region Growing Based Segmentation Algorithm for Typewritten and Handwritten Text Recognition. Applied Soft Computing 9(2), 608–617 (2009)
[12] Monjel, M.S., Khan, M.: Optical Character Recognition for Bangla Documents Using HMM. Technical Report, Centre for Research on Bangla Language Processing, CRBLP (2007)
[13] Das, D., Yasmin, R.: Segmentation and Recognition of Unconstrained Bangla Handwritten Numeral. Asian Journal of Information Technology 5(2), 155–159 (2006)
[14] Bhattacharya, U., Shridhar, M., Parui, S.K.: On Recognition of Handwritten Bangla Characters. In: Kalra, P.K., Peleg, S. (eds.) ICVGIP 2006. LNCS, vol. 4338, pp. 817–828. Springer, Heidelberg (2006)
[15] Chaudhuri, B.B.: A Complete Handwritten Numeral Database of Bangla – A Major Indic Script. In: 10th International Workshop on Frontiers in Handwriting Recognition (2006)
[16] Hasan, M.A.M., Alim, M.A., Islam, M.W.: A New Approach to Bangla Text Extraction and Recognition from Textual Image. In: 8th International Conference on Computer and Information Technology, ICCIT (2005)
[17] Mahmud, J.U., Raihan, M.F., Rahman, C.M.: A Complete OCR System for continuous Bengali Character. In: TENCON 2003, Conference on Convergent Technologies for Asia-Pacific Region, October 15-17 (2003)
[18] Bhattacharya, U., Das, T.K., Datta, A., Parui, S.K., Chaudhuri, B.B.: Recognition of Handprinted Bangla Numerals Using Neural Network Models. In: Pal, N.R., Sugeno, M. (eds.) AFSS 2002. LNCS (LNAI), vol. 2275, pp. 228–235. Springer, Heidelberg (2002)
[19] Syeed, M.M.M., Siddiqui, F.H., Al-Mamun, A.S.A., Tanbeer, S.K., Mottalib, M.A.: Bengali Character Recognition using Bidirectional Associative Memories (BAM) Neural Network. In: 5th International Conference on Computer and Information Technology (ICCIT), pp. 247–251 (2002)
[20] Karim, M.E., Hossain, M., Mottaliband, M.A., Zaman, T.: A Modified Neural Network Learning Approach and its Application to Bengali Character Recognition. Malaysian Journal of Computer Science 11(2), 68–73 (1998)
[21] Chaudhuri, B.B., Pal, U.: A Complete Printed Bangla OCR System. Pattern Recognition 31, 531–549 (1997); Graphics and Image Processing, NCCIS
[22] Dutta, A., Chaudhuri, S.: Bengali Alpha-numeric Character Recognition using Curvature Features. Pattern Recognition 26, 1757–1770 (1993)
[23] Datta, S., Chaudhury, S., Parthasarathy, G.: On Recognition of Bengali Numerals with Back Propagation Learning. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 94–99 (October 1992)
[24] Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intell. Data Anal. 11(5), 561–580 (2007)
[25] Saeed, K., Tabędzki, M., Rybnik, M., Adamski, M.: K3M – A Universal Algorithm for Image Skeletonization and a Review of Thinning Techniques. International Journal of Applied Mathematics and Computer Science 20(2), 317–335 (2010)